Requested Patent: JP2001344246A

Title:

METHOD FOR PREPARING TERM TABLE DATA BASE AND METHOD FOR RETRIEVING ELECTRONIC DOCUMENT ;

Abstracted Patent: JP2001344246 ;

Publication Date: 2001-12-14 ;

Inventor(s): IWAMOTO HAJIME ;

Applicant(s): KANSAI ELECTRIC POWER CO INC:THE ;

Application Number: JP20000160950 20000530 ;

Priority Number(s): ;

IPC Classification: G06F17/30 ;

Equivalents: ;

ABSTRACT:

PROBLEM TO BE SOLVED: To provide a method for reducing the redundancy of a retrieved result due to the ambiguity of meaning of a retrieving keyword and the variety of meaning relation among plural retrieving keywords in the retrieval of an electronic document and realizing a highly efficient electronic document retrieval.SOLUTION: The network address of an electronic document file group stored in a computer connected to a computer network is specified, a sentence including a specific term is extracted from the contents of the electronic document file group on the basis of a previously prepared extraction rule and an explanation category indicating which explanation document for terms corresponds to the extracted sentence, the network address of an electronic document file in which the extracted terms appears, the term, sentence including the term, and the type of the electronic document file are registered in a term table data base as a data set of five items.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－344246
（P2001－344246A）
(43)公開日　平成13年12月14日(2001.12.14)

(51)Int.Cl.⁷	識別記号	FI	テーマコード(参考)
G06F 17/30	170	G06F 17/30	170A 5B075
	110		110F
	230		230Z

審査請求　未請求　請求項の数7　OL　（全 7 頁）

(21)出願番号　特願2000－160950(P2000－160950)

(22)出願日　平成12年５月30日(2000.5.30)

特許法第30条第１項適用申請有り　平成12年４月10日
関西電力株式会社総合技術研究所発行の「R&D News Kansai　4月号」に発表

(71)出願人　000156938
関西電力株式会社
大阪府大阪市北区中之島３丁目３番22号
(72)発明者　岩本　元
大阪府大阪市北区中之島３丁目３番22号
関西電力株式会社内
(74)代理人　100105223
弁理士　岡崎　謙秀　（外１名）

Ｆターム(参考)　5B075 ND03 NK02 NK46 NR06 PP23
PQ02 QP03 UU01

(54)【発明の名称】　用語集データベース作成方法および電子文書検索方法

(57)【要約】
【課題】　電子文書の検索において検索キーワードの持つ意味の多義性や複数の検索キーワード間の意味的な関係の多様性による検索結果の冗長性を削減し、効率の高い電子文書検索を実現する方法を提供する。
【解決手段】　コンピュータネットワークに接続されているコンピュータに保存されている電子文書ファイル群のネットワークアドレスを指定し、これらの電子文書ファイル群の文面からあらかじめ作成されていた抽出ルールに基づき特定の用語が含まれる文章を抽出し、この抽出された文章が用語についてのいかなる解説文書であるかを示す解説カテゴリ、抽出された用語が掲載されている電子文書ファイルのネットワークアドレス、用語、および用語が記載されている文章、電子文書ファイルのタイトルの５項目を１組のデータセットとして用語集データベースに登録する。

【特許請求の範囲】

【請求項１】　コンピュータネットワークに接続されているコンピュータに保存されている電子文書ファイル群のネットワークアドレスを指定し、これらの電子文書ファイル群の文面からあらかじめ作成されていた抽出ルールに基づき特定の用語が含まれる文章を抽出し、この抽出された文章が用語についてのいかなる解説文書であるかを示す解説カテゴリ、抽出された用語が掲載されている電子文書ファイルのネットワークアドレス、用語、および用語が記載されている文章、電子文書ファイルのタイトルの５項目を１組のデータセットとして用語集データベースに登録することを特徴とする用語集データベース作成方法。

【請求項２】　請求項１記載の用語集データベース作成方法により構築された用語集データベースを、利用者により入力された検索条件に基づく検索を実行し、登録されている用語データが検索条件に合致する場合には、用語データが属するデータセットを検索結果として出力し、利用者が出力された検索結果の中から選択したデータセットに含まれるネットワークアドレスデータに基づき電子文書のネットワークアドレスを参照し、参照されたネットワークアドレスに格納されている電子文書の内容を出力することを特徴とする電子文書検索方法。

【請求項３】　請求項１記載の用語集データベース作成方法により構築された用語集データベースと、電子文書検索クライアントから送信されるデータを受信するデータ受信機能と、電子文書検索クライアントから受信した検索条件データに基づき用語集データベースに格納されたネットワークアドレスデータを対象とする検索を実行する用語集データベース検索機能と、電子文書検索クライアントから受信したネットワークアドレスデータに基づき電子文書のネットワークアドレスを参照し、参照されたネットワークアドレスに格納されている電子文書データを受信する電子文書受信機能と、用語集データベース検索機能により得られた検索結果データおよび電子文書受信機能により得られた電子文書データを電子文書検索クライアントへ送信するためのデータ送信機能とを具備する電子文書検索サーバと、利用者が外部入力装置により入力した検索条件に関するデータおよび利用者が外部入力装置により選択入力したネットワークアドレスデータを電子文書検索サーバに送信するデータ送信機能と、電子文書検索サーバより送信された検索結果データおよび電子文書データを受信するデータ受信機能と、この受信したデータを外部出力装置へ出力するデータ外部出力機能とを具備する電子文書検索クライアントとが、コンピュータネットワークにより接続されていることを特徴とする電子文書検索システム。

【請求項４】　コンピュータネットワークがインターネットであり、電子文書検索サーバがWWWサーバ機能を具備することを特徴とする請求項３記載の電子文書検索システム。

【請求項５】　検索対象である電子文書がＨＴＭＬ文書であることを特徴とする請求項３記載の電子文書検索システム。

【請求項６】　電子文書検索クライアントがWWWブラウザ機能を具備することを特徴とする請求項３記載の電子文書検索システム。

【請求項７】　請求項１記載の用語集データベース作成方法および請求項２記載の電子文書検索方法をコンピュータにより機能させるプログラムとして記録されていることを特徴とするコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この出願の発明は、用語集データベース作成方法および電子文書検索方法に関するものである。さらに詳しくは、この発明は、検索条件となる単語と電子文書中の特定の単語との意味的な関係を電子文書から抽出することで作成された検索条件となる用語に関する用語集データベースを構築する方法と、その用語集データベースをインターフェースとして利用することで実現される電子文書検索方法に関するものである。

【０００２】

【従来の技術とその課題】高速コンピュータネットワークの普及、拡大に伴い、電子化された文書情報を検索するための検索サービスが数多く提供されている。それらの多くは、検索キーワードのマッチングによる全文検索によるものであるが、次の挙げるような理由により、利用者の満足のいくような検索結果が得られない場合も多い。

【０００３】例えば、「ネットワーク」という単語は、「通信用のネットワーク」といった意味や「テレビ局やラジオ局の番組供給網」といった意味などの複数の意味を持っている。このため、利用者が「通信用のネットワーク」という意味を意図して「ネットワーク」という単語を検索キーワードとして入力した場合にも、「通信用のネットワーク」という意味以外の文書も検索結果として出力されることになり、ユーザーはその膨大な出力結果から欲する文書を選択しなければならない。

【０００４】検索キーワードの多義性に起因する問題を解決する方法として、検索キーワードを複数指定する方法がある。上記の例においては、「ネットワーク」という検索キーワードに「通信」という単語を検索キーワードとして付加し、ふたつの検索キーワードの論理積（ＡＮＤ）を検索条件として設定することで、意味が限定され検索結果もある程度絞り込まれることが期待される。しかしながら、文中における検索キーワード同士の関係が多様性を持つことから、例えば「ネットワーク」ＡＮＤ「通信」を検索条件とした場合には、結果として「通

信用のネットワーク」だけでなく、「通信会社間のネットワーク」などといったユーザーが意図した対象とは異なる文書までもが検索結果として出力される場合も少なくない。

【0005】検索キーワードの持つ意味の範囲を拡張したり類義語を得る方法として予めシソーラスデータベースを構築しておく方法が知られている。しかしシソーラスデータベースの構築は人手によるものであり、専門領域ひとつのデータベースを構築する作業であってもその作業量は膨大である。また、文書群から得られた単語の共起関係に基づき関連語をデータベース化しておき検索結果として提示する方法も提案されているが、実際の単語間の意味における類義性が低く、かえって、検索結果が冗長なものとなることも頻繁に発生する。

【0006】この出願の発明は、以上の通りの事情に鑑みてなされたものであり、電子文書の検索において検索キーワードの持つ意味の多義性や複数の検索キーワード間の意味的な関係の多様性による検索結果の冗長性を削減し、効率の高い電子文書検索を実現する方法を提供することを課題としている。

【0007】

【課題を解決するための手段】この出願の発明は、上記の課題を解決するものとして、コンピュータネットワークに接続されているコンピュータに保存されている電子文書ファイル群のネットワークアドレスを指定し、これらの電子文書ファイル群の文面からあらかじめ作成されていた抽出ルールに基づき特定の用語が含まれる文章を抽出し、この抽出された文章が用語についてのいかなる解説文書であるかを示す解説カテゴリ、抽出された用語が掲載されている電子文書ファイルのネットワークアドレス、用語、および用語が記載されている文章、電子文書ファイルのタイトルの5項目を1組のデータセットとして用語集データベースに登録することを特徴とする用語集データベース作成方法を提供する。

【0008】また、この出願の発明は、上記の用語集データベース作成方法により構築された用語集データベースを、利用者により入力された検索条件に基づく検索を実行し、登録されている用語データが検索条件に合致する場合には、用語データが属するデータセットを検索結果として出力し、利用者が出力された検索結果の中から選択したデータセットに含まれるネットワークアドレスデータに基づき電子文書のネットワークアドレスを参照し、参照されたネットワークアドレスに格納されている電子文書の内容を出力することを特徴とする電子文書検索方法を提供する。

【0009】さらに、この出願の発明は、前記の用語集データベース作成方法により構築された用語集データベースと、電子文書検索クライアントから送信されるデータを受信するデータ受信機能と、電子文書検索クライアントから受信した検索条件データに基づき用語集データベースに格納されたネットワークアドレスデータを対象とする検索を実行する用語集データベース検索機能と、電子文書検索クライアントから受信したネットワークアドレスデータに基づき電子文書のネットワークアドレスを参照し、参照されたネットワークアドレスに格納されている電子文書データを受信する電子文書受信機能と、用語集データベース検索機能により得られた検索結果データおよび電子文書受信機能により得られた電子文書データを電子文書検索クライアントへ送信するためのデータ送信機能とを具備する電子文書検索サーバと、利用者が外部入力装置により入力した検索条件に関するデータおよび利用者が外部入力装置により選択入力したネットワークアドレスデータを電子文書検索サーバに送信するデータ送信機能と、電子文書検索サーバより送信された検索結果データおよび電子文書データを受信するデータ受信機能と、この受信したデータを外部出力装置へ出力するデータ外部出力機能とを具備する電子文書検索クライアントとが、コンピュータネットワークにより接続されていることを特徴とする電子文書検索システムを提供する。この電子文書検索システムは、コンピュータネットワークがインターネットであり、電子文書検索サーバがWWWサーバ機能を具備すること、検索対象である電子文書がHTML文書であること、および、電子文書検索クライアントがWWWブラウザ機能を具備することを特徴とする。

【0010】そして、この出願の発明は、前記用語集データベース作成方法および前記電子文書検索方法をコンピュータにより機能させるプログラムとして記録されていることを特徴とするコンピュータ読み取り可能な記憶媒体をも提供する。

【0011】

【発明の実施の形態】この出願の発明は上記のとおりの特徴をもつものであるが、以下にその実施の形態について説明する。

【0012】この出願の発明の電子文書検索方法を実現する電子文書検索システムは、コンピュータネットワークに接続された用語集データベース、電子文書検索サーバ、および、電子文書検索クライアントを、基本構成とする。

【0013】用語集データベースには、用語と電子文書ファイルの文面との関係がデータとして登録される。すなわち電子文書ファイルに記載されている文面の内容が、検索キーワードに関するどのような情報であるかということが、データとして登録される。この出願の発明に係る電子文書検索は、この用語集データベースをインターフェースとして実行されるものであり、この用語集データベースは、前もって構築されていることが前提となる。用語集データベースの構築の手順について、図1を用いながら以下に示す。

【0014】まず、電子文書読み出し機能を備えるコン

ピュータ（101）によりネットワークアドレス一覧ファイル（102）が参照される。このネットワークアドレス一覧ファイル（102）に記述されているネットワークアドレスが参照され、コンピュータネットワーク（103）に接続されたコンピュータに格納されている電子文書ファイル群（104）中から、ネットワークアドレス一覧ファイル（102）により指定されたネットワークアドレスに格納されている電子文書ファイルが読み込まれ補助記憶装置（105）に登録される。次いで、用語抽出機能を備えるコンピュータ（106）により補助記憶装置（105）に保存された文書ファイル群が読み出され、あらかじめ作成されていた用語抽出ルールファイル（107）に記述されている用語抽出ルールが適用され、検索キーワードになりうる用語の抽出が行われる。用語の抽出は、電子文書の字面上のパターン分析を利用した汎用情報抽出ソフトウェアにより自動的に行われる。

【0015】用語抽出ルールは、文章が用語に関して何らかの解説をおこなっている場合においてのみ抽出が行われるように、例えば、以下の様に設定されている。

（1）漢字または英数字からなる文字列の直後に“（”、“）”（カッコ）で囲まれている文字列を抽出する。

（2）“「A」とは「B」である。”という表現から、AとBにあたる文字列を抽出する。

（3）“「A」を開発する。”という表現から、Aにあたる文字列を抽出する。

【0016】上に例示したような用語抽出ルールは用いたとき、文書ファイル中に抽出された文章がある場合には、その文章は特定の用語に関する解説を含んでいるものと判断され、抽出された用語、抽出対象となった文書ファイルのネットワークアドレス、文章が用語に関するどのような解説文であるかを表す解説カテゴリ、電子文書ファイルのタイトル、および抽出された文章の5つ項目が、1単位のデータセットとして用語集データベース（108）に登録される。ここで、解説カテゴリとしては、単語に関する「定義」、「訳語」、「性質」、「同意語」、「反意語」などが基本項目として設定されており、さらには、その用語のもつ属性に関しても各種の解説カテゴリが適宜に追加設定される。

【0017】具体的には、上記の用語抽出ルール（2）においては、「B」は「A」の「定義」の基本項目に属するものと判定される。用語抽出ルールは、解説カテゴリのそれぞれについて、予め用意されている。

【0018】この出願の発明に係る電子文書検索方法について、図2～6を用いて説明する。

【0019】まず、電子文書検索サービス利用者は、電子文書検索クライアント（201）の外部入力装置（202）により検索条件を入力する。入力された検索条件は、電子文書検索クライアント（201）の備えるデータ送信機能により電子文書検索サーバ（203）に送信される。

【0020】電子文書検索サーバ（203）においては、受信した検索条件に基づき用語集データベース（204）に格納されている用語データを検索し、検索条件と一致するものがあれば、その用語データに関するデータセット（用語データ、組となる文書ファイルのネットワークアドレス、解説カテゴリ、文章、電子文書ファイルのタイトル）は全て、検索結果として電子文書検索サーバ（203）の備え持つデータ送信機能により電子文書検索クライアント（201）へと送信される。

【0021】電子文書検索クライアント（201）においては、受信した検索結果であるデータセット（用語データ、文書ファイルのアドレスデータ、解説カテゴリデータ、文書データ）が外部出力装置（205）により出力される。

【0022】外部出力装置に出力される内容については、図3に例示するように、まず、検索条件（301）が用語集データベースに登録された用語にヒットした場合には、表示欄（A）にヒットした用語（302）が表示される。また、検索条件（301）が、用語集データベースに登録された用語を解説する文書中に記載されていた場合には、表示欄（B）に、解説対象である用語（303）が表示される。

【0023】電子文書検索サービス利用者が、電子文書検索クライアントに接続されている外部入力装置を用いて、表示されたヒットした用語（302）の中から目的とする用語を選択すると、図4に例示したように、その用語に対応する用語データの解説カテゴリ（401）の一覧が表示される。次いで、電子文書検索サービス利用者が、電子文書検索クライアントに接続されている外部入力装置を用いて、表示された解説カテゴリ（401）の中から目的とする解説カテゴリを選択すると、図5に例示したように、表示欄（C）に選択した解説カテゴリに関する電子文書ファイルの文章（501）と電子文書ファイルのタイトル（502）が表示される。

【0024】さらに、電子文書検索サービス利用者が、電子文書検索クライアントに接続されている外部入力装置を用いて、表示された電子文書ファイルのタイトル（502）の中から目的とする電子文書ファイルのタイトルを選択することで、選択に関するデータが電子文書検索サーバに送信され、電子文書検索サーバは選択された電子文書ファイルを読み出し、電子文書検索クライアントに送信する。具体的には、図6に示すように、電子文書検索クライアント（601）においては、電子文書検索サービス利用者が外部入力装置（602）より選択した文書ファイルに関するネットワークアドレスデータが、電子文書検索クライアント（601）の備えるデータ送信機能により電子文書検索サーバ（603）へと送信される。

【0025】電子文書検索サーバ（603）においては、電子文書検索サービス利用者により選択した文書ファイルのネットワークアドレスデータに基づき、電子文書検索サービス利用者により選択された文書ファイルの格納されているコンピュータネットワーク（604）に接続されたコンピュータから電子文書ファイル（605）が読み込まれる。読み込まれた電子文書ファイルは、電子文書検索クライアント（601）へ送信され、文面が外部出力装置（605）により外部出力される。このとき、外部出力装置には、電子文書検索サービス利用者が図5において選択した用語の解説文にあたる部分が出力画面の最初に来るように、自動的に表示される。

【0026】この出願において、コンピュータネットワークは、どのような規模を持つものであってもよく、また、ネットワークプロトコルやネットワークトポロジーに関しても、特に限定されるものではない。例えば、コンピュータネットワークはＴＣＰ／ＩＰにより接続された企業内ＬＡＮやインターネットであり、電子文書検索サーバはWWWサーバとしての機能を備え、また検索対象である電子文書がＨＴＭＬ文書である。このとき、電子文書検索クライアントは、WWWブラウザとしての機能を具備するものである。

【0027】さらに、この出願の発明の用語集データベース作成方法および電子文書検索方法は、コンピュータにより機能させるプログラムとして記憶媒体に記録される。

【0028】

【発明の効果】以上、詳しく説明した通り、この出願の発明により、電子文書の検索において検索キーワードの持つ意味の多義性や複数の検索キーワード間の意味的な関係の多様性による検索結果の冗長性を削減し、効率の高い電子文書検索方法が提供される。この出願の発明により、検索キーワードの多義性を意識した検索が可能となり、検索キーワードに関連する情報をも含む情報に対して系統立てられた検索が実現する。

【図面の簡単な説明】

【図1】この出願の発明である用語集データベース作成方法における処理とデータの流れを示した概要図である。

【図2】この出願の発明である電子文書検索方法における処理とデータの流れを示した概要図である。

【図3】この出願の発明である電子文書検索方法において電子文書検索クライアントの外部出力装置に出力される画面構成を例示した概要図である。

【図4】この出願の発明である電子文書検索方法において電子文書検索クライアントの外部出力装置に出力される画面構成を例示した概要図である。

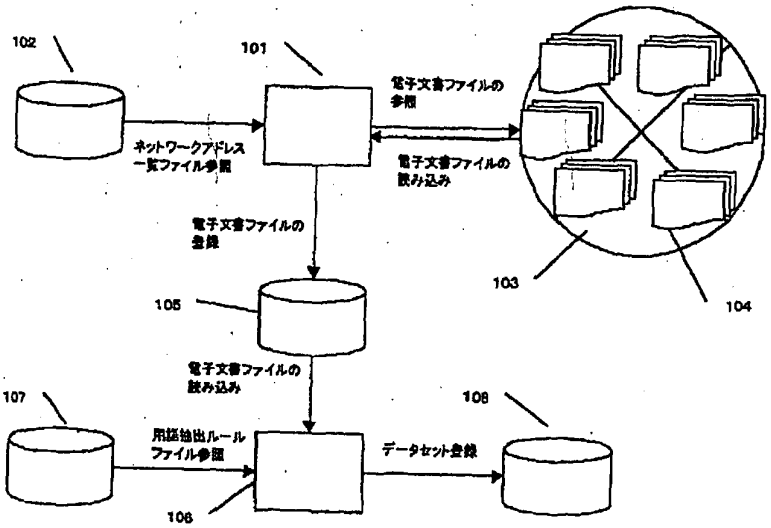
【図5】この出願の発明である電子文書検索方法において電子文書検索クライアントの外部出力装置に出力される画面構成を例示した概要図である。

【図6】この出願の発明である電子文書検索方法における処理とデータの流れを示した概要図である。

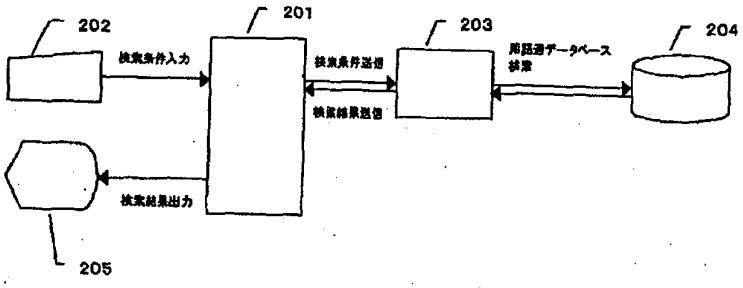
【符号の説明】

101　電子文書読み出し機能を備えるコンピュータ
102　ネットワークアドレス一覧ファイル
103　コンピュータネットワーク
104　電子文書ファイル群
105　補助記憶装置
106　用語抽出機能を備えるコンピュータ
107　用語抽出ルールファイル
108　用語集データベース
201　電子文書検索クライアント
202　外部入力装置
203　電子文書検索サーバ
204　用語集データベース
204　電子文書検索サーバ
205　外部出力装置
301　検索条件
302　用語
303　用語
401　解説カテゴリ
501　電子文書ファイルの文章
502　電子文書ファイルのタイトル
601　電子文書検索クライアント
602　外部入力装置
603　電子文書検索サーバ
604　コンピュータネットワーク
605　電子文書ファイル
606　外部出力装置

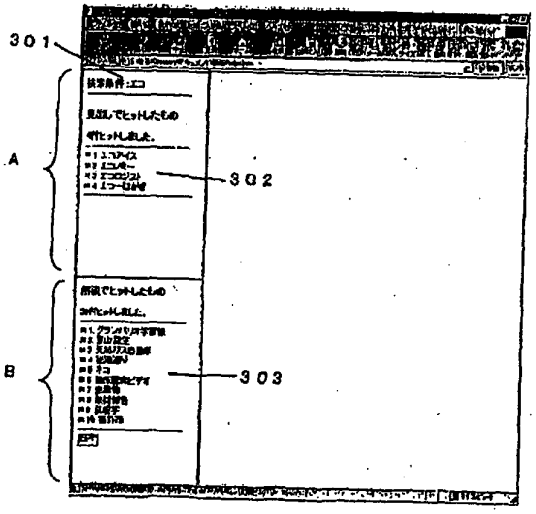
【図1】


102
101
電子文書ファイルの参照
ネットワークアドレス一覧ファイル参照
電子文書ファイルの読み込み
電子文書ファイルの登録
103
104
105
電子文書ファイルの読み込み
107
108
用語抽出ルールファイル参照
データセット登録
106

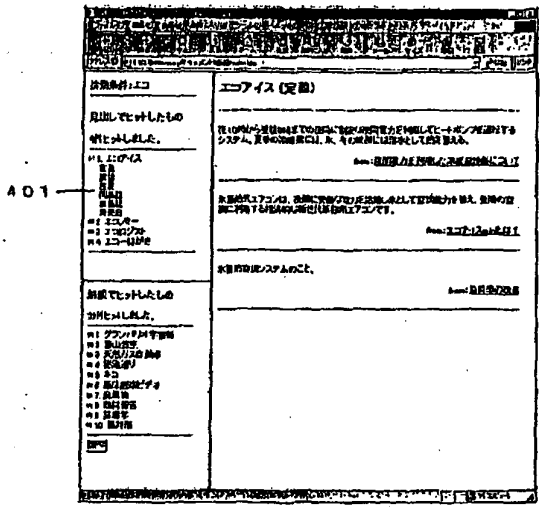
【図2】


202
201
203
204
検索条件入力
検索条件送信
検索結果送信
用語辞データベース検索
検索結果出力
205

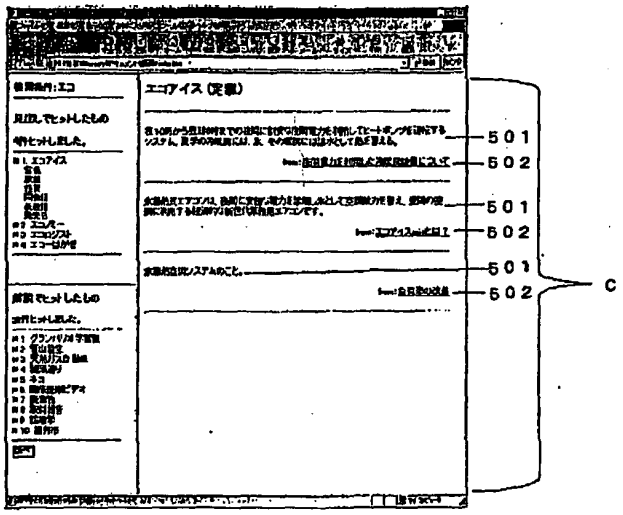
【図3】


301
A
302
B
303

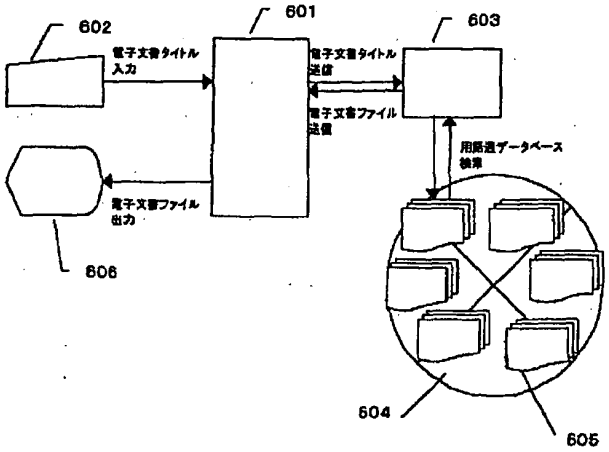
【図4】


401

【図5】


501
502
501
502
501
502
C

【図6】


602
601
603
電子文書タイトル
入力
電子文書タイトル
送信
電子文書ファイル
送信
用語選データベース
検索
電子文書ファイル
出力
606
604
605